

ABSTRACT

Congestion within a communication is controlled by rate limiting packet transmissions over selected communication links within the network and modulating the rate limiting according to buffer occupancies at control nodes within the network. Preferably, though not necessarily, the rate limiting of the packet transmissions is performed at an aggregate level for all traffic streams utilizing the selected communication links. The rate limiting may also be performed dynamically in response to measured network performance metrics; such as the throughput of the selected communication links input to the control points and/or the buffer occupancy level at the control points. The network performance metrics may be measured according to at least one of: a moving average of the measured quantity, a standard average of the measured quantity, or another filtered average of the measured quantity. The rate limiting may be achieved by varying an inter-packet delay time over the selected communication links at the control points. The control points themselves may be located upstream or even downstream (or both) of congested nodes within the network and need only be located on only a few of a number of communication links that are coupled to a congested node within the network. More generally, the control points need only be associated with a fraction of the total number of traffic streams applied to a congested node within the network.